

Exploring probabilistic grammar(s) in varieties of English around the world

**Grafmiller, Jason; Heller, Benedikt;
Röthlisberger, Melanie; Szmrecsanyi,
Benedikt**

University of Leuven, Belgium

We report on an ongoing project marrying the probabilistic grammar framework (Bresnan et al. 2007; Bod, Hay and Jannedy 2003) to research along the lines of the "English World-Wide Paradigm", which is concerned with the sociolinguistics of, and linguistic variation across, post-colonial English-speaking communities around the world (Mesthrie 2006; Schneider 2007). Through state-of-the-art quantitative analyses of syntactic alternations across across nine varieties of English represented in the International Corpus of English (St BrE, Hong Kong E, Jamaican E, Canadian E, Singapore E, Philippines E, Indian E, New Zealand E, Irish E), the project addresses the central question: *How does language users' grammatical knowledge differ across post-colonial varieties of English?* In treating variation as a "core explanandum" (Adger and Trousdale 2007: 274) of linguistic theory, the project contributes to the development of usage-based theoretical linguistics by adopting a variational, large-scale comparative, and sociolinguistic perspective.

Previous research within the English World-Wide Paradigm has primarily focused on the presence, absence, or variable usage frequencies of grammatical features (e.g. Kortmann and

Szmrecsanyi 2004; Steger and Schneider 2012). While such studies are undoubtedly interesting, this project is more ambitious in scope. It is an investigation of linguistic knowledge that goes beyond mere description of frequencies, and links the EnglishWorld-Wide Paradigm to recent advances in usage- and experience-based models of language (see also Bernaisch et al. To appear). The project explores variability in the hidden— though cognitively 'real'— probabilistic constraints that fuel variation within and across speech communities. Such constraints, e.g. the tendency to place long constituents after short constituents (Wasow and Arnold 2003), are not necessarily tied to surface material but to subtle stochastic generalizations about language usage, which— according to experimental evidence (Ford and Bresnan 2010)— language users implicitly know about. Thus, rather than simply describing probabilistic variation in corpus data, the project aims to illuminate aspects of the linguistic *knowledge* that language users with differing English backgrounds implicitly command.

We accomplish this objective by investigating the probabilistic effects of various constraints on linguistic choice-making as a function of regional variety using advanced statistical methods, primarily mixed-effects logistic regression modeling (Pinheiro and Bates 2000) and conditional random forests analysis (Tagliamonte and Baayen 2012). Four patterns of grammatical variation ("alternations") sensitive to a range of constraints take center stage: the GENITIVE ALTERNATION, the DATIVE ALTERNATION, PARTICLE PLACEMENT, and (NON-)FINITE COMPLEMENTATION. Each of these alternations are examined within and across data from parallel spoken and written registers in each of the nine English varieties. Throughout this endeavor, we focus on questions related to, among other things: the extent to which different English varieties share a core grammar that is explanatory across different varieties; the ways implicit knowledge of linguistic probabilities develops in historically diverging groups of speakers; the degree to which individual probabilistic

constraints are cross-lectally, sociohistorically, and/or culturally malleable; and the degree to which the alternations under study exhibit cross-constructural parallelisms.

References can be provided upon request